한국음향학회지 제21권 제1호 pp. 11~18 (2002)

잡음하에서 이득 적응을 가지는 비정상상태 자기회귀 은닉 마코프 모델에 의한 오염된 음성을 위한 인식

Recognition for Noisy Speech by a Nonstationary AR HMM with Gain Adaptation Under Unknown Noise

서 창 우*, 이 주 헌**, 이 기 용*
(Changwoo Seo*, Joohun Lee**, Ki Yong Lee*)

*숭실대학교 정보통신전자공학부, **동아방송대학 인터넷방송과
(접수일자: 2001년 8월 29일; 수정일자: 2001년 11월 9일; 채택일자: 2001년 12월 10일)

본 논문에서는 부가 잡음에 오염된 음성신호에 이득 적응을 가지는 음성인식을 시간 영역에서 다루었다. 잡음은 유색잡음이라고 가정한다. 전화망에서 마찰음 (fricative), 운음 (glides), 유음 (liquds), 그리고 천이영역 (transition region)과 같은 음성 신호의 뚜렷한 비정상상태를 극복하기 위해서 NAR-HMM (nonstationary autoregressive HMM)을 제안하였다. 비정상상태 AR 처리는 M개의 알고 있는 기저 함수 (basis function)의 선형 결합으로 이루어진 다항 함수 (polynomial function)로 나타낼 수 있다. 오염된 신호만을 이용할 수 있을 때, 잡음의 추정 (estimation) 문제는 필연적으로 발생한다. 다중 Kalman 필터를 사용함으로써, 잡음모델의 추정과 음성의 이득 곡선 (gain contour)을 수행하였다. 제안한 방법의 잡음 추정은 오염된 신호로부터 효과적으로 잡음을 제거하여 깨끗한 음성신호를 얻을 수 있었다. 또한 잡음 추정을 하는 일반적인 ARHMM보다 제안한 NAR-HMM이 약 2-3%의 인식성능을 향상시켰다.
핵심용어: 비선형 자기회귀-은닉마코프모델, 다중 칼만 필터, EM 알고리즘, 음성향상, 음성인식
투고분야: 음성처리 분야 (2.5)

In this paper, a gain-adapted speech recognition method in noise is developed in the time domain. Noise is assumed to be colored. To cope with the notable nonstationary nature of speech signals such as fricative, glides, liquids, and transition region between phones, the nonstationary autoregressive (NAR) hidden Markov model (HMM) is used. The nonstationary AR process is represented by using polynomial functions with a linear combination of M known basis functions. When only noisy signals are available, the estimation problem of noise inevitably arises. By using multiple Kalman filters, the estimation of noise model and gain contour of speech is performed. Noise estimation of the proposed method can eliminate noise from noisy speech to get an enhanced speech signal. Compared to the conventional ARHMM with noise estimation, our proposed NAR-HMM with noise estimation improves the recognition performance about 2-3%.
Keywords: *NAR-HMM, Multiple Kalman filters, EM algorithm, Speech enhancement, Speech recognition*
ASK subject classification: *Speech signal processing (2.5)*

책임저자: 이기용 (kylee@saint.soongsil.ac.kr)
156-743 서울시 동작구 상도동
숭실대학교 정보통신전자공학부
(전화: 02-820-0908; 팩스: 02-821-7653)

I. 서 론

음성인식의 초창기부터, 인식 시스템의 성능 감쇠는 시스템이 열악한 조건에서 사용될 때 나타나는 큰 문제들 중의 하나였다. 음성인식 시스템의 실제 응용은 배경잡음 (background noise), 채널 간섭 (channel interference) 그리고 마이크 왜곡 (microphone distortion)과 같은 실제 조건의 많은 다양성 (diversity) 때문에 성능 감쇠는 필연적이다. 특히 그런 여러가지 장애는 학습 과정과 테스트 과정의 조건이 다르기 때문에 주로 발생한다. 높은 인식율을 유지하기 위해서는 입력 음성의 질이 떨어질 때나 학습과 테스트 환경의 음성 특성이 다를 때에도 강인한 음성인식이 필요하다. 열악한 조건에서 일어나는 장애에 대해서 더 좋은 성능을 얻기 위해서 음성인식 시스템이 연구되어 왔다[1].

ARHMM (autoregressive HMM)[2,3]은 음성인식과 향상에서 깨끗한 음성을 얻기 위한 좋은 방법이다. 일반적인 ARHMM에서 각각의 상태는 정상상태 통계열이라고 가정한다. 전화망에서 마찰음, 운음, 유음 그리고 천이 영역과 같은 음성은 가장 두드러진 비정상상태 특성을 나타내기 때문에[4-6], 위의 가정에 기초를 둔 일반적인 방법으로 좋은 성능을 얻는 것을 기대하기란 어렵다. 이런 문제점을 극복하기 위해서 다항 회귀 함수를 이용한 HMM이 제안되었다[7-9]. 이러한 방법은 파라미터 영역에 기초를 두고 문제에 접근하지만, 논문에서 제안하는 방법은 잡음 추정에 Kalman 필터를 사용함으로써 시간 영역에 기초를 둔 이득 적응을 하는 음성인식 모델을 적용하였다.

부가 잡음에 오염된 음성신호만이 주어질 때, ARHMM의 또 다른 문제가 음성인식에서 발생하였다. 그것은 신호를 모델링하기 위한 잡음 추정 문제와 신호의 에너지 곡선을 매칭시키는 문제이다. 음성인식에서 PMC (Parallel Model Combination)[10]와 같은 방법이 부가 잡음과 컨벌루션 (convolutional) 잡음에 제안되었다. PMC는 열악한 잡음에 대해서 좋은 인식 성능을 보여준다. 그러나 이러한 방법은 음성의 비정상상태 특성을 반영하지는 않는다.

본 논문에서는 NAR-HMM의 이득 적응된 음성인식과 잡음 추정을 소개한다. NAR-HMM은 깨끗한 음성을 모델링하기 위해서 사용되었으며, 그리고 비정상상태 AR 모델의 파라미터는 M개의 알고 있는 기저 함수의 선형 조합으로 이루어졌다. 이때 음성 신호는 고정된 길이의 프레임 단위로 블록화시켰다. 제안한 모델은 trend HMM[4,5]과 매우 비슷하지만, 특징 벡터 (feature vectors)를 직접 다루기 보다는 신호에 의한 프레임 레벨에서의 음성 신호를 다루도록 설계되었다. 또한 M=0일 때, 제안한 모델은 일반적인 ARHMM[3]이 된다. 잡음이 부가된 신호가 있을 때, 잡음 추정을 하는 이득 적응된 인식 알고리즘은 EM 접근을 사용하여 NAR-HMM을 발전시켰으며, 잡음이 존재하는 음성인식에 테스트하였다. 또한 다중 Kalman 필터를 사용해서 잡음 모델을 추정하고 음성의 이득 곡선을 계산하였다.

논문은 다음과 같이 구성되었다. II장은 깨끗한 음성에 대한 비정상상태 AR-HMM을 구성하여 음성신호의 시변이 (time-varying) 계수에 대한 작용을 설명하고 있으며, 추정 문제을 다루기 쉽게 하기 위해서 M개의 알고 있는 기저 함수로 모델링하였다. III장은 NAR-HMM에 대한 이득 적응된 학습 알고리즘을 설명하였으며, 비정상상태 ARHMM의 파라미터와 음성신호에 대한 이득 곡선은 EM 알고리즘으로 추정하였다. IV장에서는 음성에 잡음이 부가되었을 때 잡음에 대한 음성인식을 설명하고 있다. 잡음을 유색잡음이라고 가정하여 AR 모델로 모델링하였다. 테스트 결과는 V장에서 설명되며, VI장에는 결론을 서술하였다.

II. 깨끗한 음성신호를 위한 NAR-HMM

음성신호 벡터 열 $y=\{\ \boldsymbol{y}_n, n=1,\ldots,T\}$라고 하자. 이때, $\boldsymbol{y}_n=\{y(t),(n-1)N+1\le t\le nN\}$이고, N과 n은 프레임 길이와 프레임 개수를 나타낸다. $g=\{g_n, n=1,\ldots,T\}$는 신호 y에 대한 이득 인자 (gain factor)의 열 또는 이득 곡선 (gain contour)이라 하자. 이때 n번째 프레임에서 상태 j에 있는 조건의 음성신호는 아래 식과 같이 과거 값을 선형조합하고 그리고 이득 곡선을 가지는 여기신호를 합한 것이다.

$$y(t)=\sum_{k=1}^{p}B_k^j(n)y(t-k)+g_n\cdot e_j(t),\qquad (n-1)N+1\le t\le nN \quad (1)$$

여기서 $B_k^j(n)$은 상태 종속 (state-dependent) 프레임 변이 계수이며, $e_j(\cdot)$는 상태 종속 분산 $\sigma_j^2(n)$을 가지는 여기신호이다. 모든 n에 있어서 인자 $g_n>0$인 것은 음성 모델에 대한 학습 데이터와 테스트 데이터의 차이 (mismatching)를 설명하기 위한 이득 항이다.

제안한 모델에서 시변이 계수의 추정문제를 설명한다. 계수의 작용을 이해하고 추정 문제를 다루기 쉽게 하기 위해서, M개의 알고있는 기저 함수의 선형 조합으로 이것을 모델링하였다.

$$B_k^i(n) = \sum_{m=0}^{M} B_{k,m}^i f_m(n) \tag{2}$$

여기서 $f_m(n)$과 $B_{k,m}^i$은 각각 m번째 기저 함수와 이 기저 함수와 관계되는 가중치 (weight)이다. 이 식은 단지 약간의 규칙적인 방법에서 계수 $B_k^i(n)$의 전개를 빠르게 하고, 스무딩한 파라미터의 전개를 빠르게 모델링할 수 있게 한다. 여기서 M=1일 때, 기저 함수는 다음과 같이 나타낼 수 있다[11].

$$\begin{aligned} f_0(n) &= 1, \qquad 1 \le n \le T \\ f_1(n) &= n, \qquad 1 \le n \le T \end{aligned} \tag{3}$$

식 (3)에서와 같이 기저 함수의 간단한 방법을 선택함으로 복잡한 방법을 도입하지 않고도 성능을 향상시킬 수 있다. 다른 방법들은 약간 더 좋은 성능을 얻을 수 있지만, 훨씬 많은 계산량을 필요로 하게 된다. 간단한 기저 함수가 주어졌을 때, M=1은 이런 경우에 적당하다. $M \ge 2$일 때, 훨씬 많은 계산량에도 불구하고 성능개선은 커지 않다[11]. 물론, 학습 데이터로 단어 (word)보다는 문장 (sentence)을 사용할 때, M=1의 선택과 간단한 기저 함수의 선택은 적당하게 변화시킬 수 있다.

따라서, 식 (1)은 (4)와 같이 벡터 형태로 다시 나타낼 수 있다.

$$\begin{aligned} y(t) = {} & \mathbf{B}^i Y(t-1) + g_n \cdot e_i(t), \\ & (n-1)N+1 \le t \le nN \end{aligned} \tag{4}$$

여기서 $\mathbf{B}^i = [B_{1,0}^i \; B_{1,1}^i \; B_{2,0}^i \; B_{2,1}^i, \cdots, B_{p,0}^i \; B_{p,1}^i]$, $Y(t-1) = [y((n-1)N+t-1), ny((n-1)N+t-1), \cdots, y((n-1)N+t-p), ny((n-1)N+t-p)]^T$이다.

실질적으로, 시변이 계수 (time-variant coefficients)를 갖는 모델은 시불변 가중치 (time-invariant weights)의 모델로 변형시킬 수 있다. 따라서 이 문제는 계수의 작용을 완벽하게 특성화시키는 $2P$개의 시불변 파라미터를 추정하는 것으로 줄어든다. 그러나 이것이 기저 함수의 선택이 다항식에 제한된다는 것을 의미하지는 않는다.

모델 λ와 이득곡선 g에서 관측열 y의 유사도는 다음과 같이 나타낼 수 있다.

$$p_\lambda(y \mid g) = \prod_{n=0}^{T} a_{s_{n-1}s_n} p_\lambda(y_n \mid s_n, g_n) \tag{5}$$

여기서 $a_{s_{n-1}s_n}$는 시간 n-1의 상태에서 시간 n의 상태로 천이확률을 나타내고, 그리고 $p\lambda(y_n \mid s_n, g_n)$은 아래 식과 같다.

$$p\lambda(y_n \mid s_n, g_n) = \prod_{t=(n-1)N}^{nN} \frac{\exp\left\{ -\dfrac{\left(y(t) - \boldsymbol{B}^{s_n} Y(t-1)\right)^2}{2g_n^2 \sigma_{s_n}^2} \right\}}{\sqrt{2\pi g_n^2 \sigma_{s_n}^2}}.$$

깨끗한 음성에 대한 비정상상태 ARHMM의 파라미터 $\lambda = \{a_{ij}, \; \boldsymbol{B}^i, \sigma_j^2, i,j = 1, \ldots, L\}$와 이득 곡선 g는 깨끗한 음성신호의 학습열로부터 추정된다. 파라미터 λ는 이득 정규화된 신호에 대한 NAR-HMM의 파라미터 집합을 나타낸다.

III. 파라미터를 얻기 위한 학습 알고리즘

비정상상태 ARHMM의 이득 적응된 학습은 학습열 y로부터 파라미터 λ와 이득 곡선의 ML (maximum likelihood) 추정의 결과이다. 이때 λ와 g는 다음과 같이 추정할 수 있다.

$$\max_{\lambda} \max_{g} p_\lambda(y \mid g) \tag{6}$$

그러나 (λ, g)에 의한 $p_\lambda(y \mid g)$의 기울기 식은 비선형이기 때문에 해결하기가 쉽지 않다. 따라서 (λ, g)의 추정은 다음과 같이 임의의 함수 (auxiliary function)를 최대화시키는 EM 접근[12,13]을 반복적으로 사용하여 얻을 수 있다.

$$Q(\lambda, g) = \sum_s p_{\lambda'}(s \mid y, g) \log p_\lambda(y, s \mid g) \tag{7}$$

여기서 λ'와 λ는 각각 모델 파라미터의 현재와 새로운 추정을 나타낸다. g가 주어질 때 λ를 추정하기 위해서 EM이 반복되고, 다음 반복에서 λ가 주어지면 EM 접근에 의한 g를 추정하는 과정이 계속해서 반복된다.

O. 파라미터 λ의 추정 (Estimation);

표준 HMM과 같이 재추정 (reestimation) 식은 다음과 같이 나타낼 수 있다.

$$a_{ij}=\frac{\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)}{\sum_{j=1}^{I}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)} \tag{8}$$

$$B^j=\Big[\sum_{i=1}^{I}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)\sum_{t=(n-1)N+1}^{nN} \overline{Y}(t-1)\,\overline{Y}^T(t-1)\Big]^{-1}\cdot\Big[\sum_{i=1}^{I}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)\sum_{t=(n-1)N+1}^{nN}\overline{y}(t)\,\overline{Y}(t-1)\Big] \tag{9}$$

$$\sigma_j^2=\frac{\sum_{i=1}^{I}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)\sum_{t=(n-1)N+1}^{nN}(\overline{y}(t)-B^j\overline{Y}(t-1))^2}{\sum_{i=1}^{I}\sum_{n=1}^{T} p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)} \tag{10}$$

여기서 $\overline{y}(t)=\frac{y(t)}{g_n}$ 이고,

$\overline{Y}(t-1)=\Big[\frac{y(t-1)}{g_n}\ n\frac{y(t-1)}{g_n}\ \frac{y(t-2)}{g_n}\ n\frac{y(t-2)}{g_n}\ \ldots\frac{y(t-p)}{g_n}\ n\frac{y(t-p)}{g_n}\Big]^T$ 이다.

확률 $p_{\lambda'}(s_{n-1}=i, s_n=j \mid y, g)$는 전향-후향 (forward-backward) 알고리즘에 의해 효과적으로 계산할수 있다. 만약 M이 0이면, 식 (8)-(10)은 표준 ARHMM에서 가우시안 평균 벡터를 위한 재추정 식이 된다.

O. 이득 곡선 g의 추정;

이득 곡선 g를 추정하기 위해서 모델 λ는 알고 있는 것으로 가정한다. 이득 곡선 g에 관한 임의의 함수의 최대화는 g를 0으로 두는 $Q(\lambda, g)$의 기울기 (gradient)를 계산함으로 g를 추정할 수 있다. 따라서 이득 재추정 식은 다음과 같다:

$$g_n^2=\sum_{j=1}^{I} p_{\lambda'}(s_n=j \mid y, g')\sum_{t=(n-1)N-1}^{nN}\frac{(y(t)-B^jY(t-1))^2}{\sigma_j^2} \tag{11}$$

반복 과정은 모든 n에 대해서 초기 이득 곡선 $g'_n=1$로 시작하며, 고정점 $\lambda=\lambda'$과 $g_n=g'_n$이 성립되거나 두개의 연속적인 반복에서 유사도의 차이가 아주 작을 때까지 계속해서 반복된다.

IV. 잡음에 오염된 음성의 인식

잡음이 부가된 음성 $z=y+v$을 이용할 때, $z=\{z_n, n=1,\ldots,T\}$과 $v=\{v_n, n=1,\ldots,T\}$, 음성의 이득 인자는 잡음이 있는 음성을 음성신호의 에너지 곡선과 매칭시킴으로서 같은 신호에 대한 모델의 이득 인자를 추정할 수 있다. 여기서 가정된 잡음 모델은 차수 q의 유색 통계 AR이다:

$$v(t)=C^TV(t-1)+g_v w(t) \tag{12}$$

여기서 $V(t-1)=[v(t-1),\ldots,v(t-q)]^T$, $C=[c_1,\ldots,c_q]^T$는 잡음처리의 AR 파라미터 벡터, g_v는 전력 (power), $w(t)$는 평균 (mean)이 0이고 분산 (variance)이 1인 백색 가우시안 (white Gaussian)이다. 잡음 파라미터 C와 전력 g_v는 사전에 알 수 없기 때문에 음성인식 알고리즘내에서 추정해야만 한다.

음성신호로부터 모델 λ가 주어지면, 잡음 모델 $\lambda_v=\{C, g_v\}$과 이득 곡선 g에서 잡음이 존재하는 음성 z의 유사도 (likelihood)는 다음과 같이 계산할 수 있다.

$$p_\lambda(z \mid g, \lambda_v)=\sum_s\int p_\lambda(s, y, z \mid g, \lambda_v)dy \tag{13}$$

여기서

$$p_\lambda(s, y, z \mid g, \lambda_v)=\prod_{n=1}^{T} a_{s_{n-1}s_n}\, p_\lambda(y_n \mid s_n, g_n, \lambda_v)\, p_{\lambda_v}(z_n-y_n)$$

이때 g와 λ_v는

$$\max_{g,\lambda_v} p_\lambda(z \mid g, \lambda_v) \tag{14}$$

로부터 추정할 수 있다. g와 λ_v에 관한 기울기 식은 비선형이기 때문에 ML 추정은 임의의 함수를 반복적으로 사용하여 얻을 수 있다.

$$Q(\lambda_v, g)=\sum_{n=1}^{T}\sum_{s_n} p_\lambda(s_n \mid z, g_n)\int p(y_n \mid g_n, z, s_n)\cdot\log p(z_n, y_n, s_n \mid g_n, \lambda_v)dy_n \tag{15}$$

식 (14)를 이용해서, 식 (15)는 다음과 같이 나타낼 수 있다.

$$\begin{aligned}Q(\lambda_v, g)&=\sum_{n=1}^{T}\sum_{s_n} p_\lambda(s_n \mid z, g_n)\int p_\lambda(y_n \mid z, s_n, g'_n)[\log a_{s_{n-1}s_n}\\&\quad+\log p_\lambda(y_n \mid s_n, g_n, \lambda_v)+\log p_{\lambda_v}(z_n-y_n)]dy_n\\&=\sum_{n=1}^{T}\sum_{s_n} p_\lambda(s_n \mid z_n, g_n)[\log a_{s_{n-1}s_n}\\&\quad+E\{\log p(y_n \mid s_n, g_n) \mid z, s_n, g'_n\}\\&\quad+E\{\log p_\lambda(z_n-y_n) \mid z, s_n, g'_n\}]\end{aligned} \tag{16}$$

상태 $s_n=j$에서, 다음과 같이 표기하였다.

$$(\hat{\cdot})_j = E\{\cdot \mid z, s_n = j, g'\} \tag{17}$$

g와 λ_v에 의한 $Q(\lambda_v, g)$의 기울기 식으로부터, 다음의 재추정 식을 얻을 수 있다:

$$g_n^2 = \sum_{j=1}^{L} p_\lambda(s_n = j \mid z, g') \sum_{t=(n-1)N+1}^{nN} \frac{\{\hat{y}_j(t) - B^j \hat{Y}_j(t-1)\}^2}{\sigma_j^2} \tag{18}$$

$$C_n = \left[\sum_{j=1}^{L} p_\lambda(s_n = j \mid z, g') \sum_{t=(n-1)N+1}^{nN} \hat{Y}_j(t-1) \hat{Y}_j^T(t-1)\right]^{-1} \cdot \left[\sum_{j=1}^{L} p_\lambda(s_n = j \mid z, g') \sum_{t=(n-1)N+1}^{nN} \hat{y}_j(t) \hat{Y}_j(t-1)\right] \tag{19}$$

$$g_{v,n}^2 = \sum_{j=1}^{L} p_\lambda(s_n = j \mid z, g') \sum_{t=(n-1)N+1}^{nN} \{\hat{v}_j(t) - C_n^T \hat{Y}_j(t-1)\}^2 \tag{20}$$

여기서 g'는 이전의 반복에서 얻어진 이득 곡선 추정이다. $p_\lambda(s_n = j \mid z, g')$는 전향-후향 과정[3]을 사용하여 효과적으로 계산할 수 있다. 프레임 n=1에서 잡음 모델 λ_v의 초기 조건은 음성이 없는 구간이 검출된 곳의 신호로부터 λ_v를 추정할 수 있다. 또한 n>1일 때, (n-1)번째 프레임에서 얻어진 λ_v를 이용하여 추정할 수 있다.

식 (18)-(20)에서, $E\{\cdot \mid z, s_n = j, g'\}$는 아래 상태-공간 (state-space) 식[14-16]과 같이 상태 $s_n = j$을 가지는 Kalman 필터로부터 얻을 수 있다.

$$X(t) = \Phi(j)X(t-1) + Gr_j(t) \tag{21}$$

$$z(t) = H^T X(t) \tag{22}$$

여기서 $X(t) = [y(t)...y(t-p)v(t)...v(t-q)]^T$,

$r_j(t) = [e_j(t)\ w(t)]^T$, $\Phi(j) = \begin{bmatrix} \Phi_y(j) & 0 \\ 0 & \Phi_v \end{bmatrix}$,

$G = \begin{bmatrix} G_y & 0 \\ 0 & G_v \end{bmatrix}$, $H^T = [H_y^T\ H_v^T]$,

$H_y^T = [10...0]$, $H_v^T = [10...0]$,

$G_y = [g_n\ 0...0]$, $G_v = [g_v\ 0...0]$,

$$\Phi_y(j) = \begin{bmatrix} B_{1,0}^j + B_{1,1}^j n & \cdots & B_{p,0}^j + Bp_{1,1}^j n & 0 \\ 0 & & I & 0 \end{bmatrix}$$

그리고 $\Phi_v = \begin{bmatrix} c_1 & \cdots & c_q & 0 \\ 0 & & I & 0 \end{bmatrix}$.

상태 $s_n = j$에서 $\hat{X}_j(t)$의 추정은 일반적인 Kalman 필터로부터 얻을 수 있다.

$$\hat{X}_j(t) = \Phi(j)\ \hat{X}_j(t-1) + K_j \cdot \{z(t) - H^T \Phi(j)\ \hat{X}_j(t-1)\} \tag{23}$$

$$M_j(t) = \Phi(j) P_j(t-1) \Phi^T(j) + GQ(j) G^T \tag{24}$$

$$K_j(t) = M_j(t) H \left[H^T M_j(t) H\right]^{-1} \tag{25}$$

$$P_j(t) = M_j(t) - K_j(t) H M_j(t) \tag{26}$$

여기서 K_j, M_j, 그리고 P_j는 각각 Kalman 이득 행렬 (gain matrix), 사전 오류 공분산 행렬 (a priori error covariance matrix), 오류 공분산 행렬 (error covariance matrix)이며, $Q(j) = E\{r_j(t) r_j^T(t)\}$이다. 식 (18)-(20)의 $\hat{y}_j^2(t)$, $\hat{Y}_j(t-1)\ \hat{Y}_j^T(t-1)$, $\hat{y}_j(t)\ \hat{Y}_j(t-1)$, $\hat{v}_j^2(t)$, $\hat{V}_j(t-1)\ \hat{V}_j^T(t-1)$ 그리고 $v_j(t)\ \hat{V}_j^T(t-1)$는 $\hat{X}_j(t)\ \hat{X}_j^T(t)$로부터 추정할 수 있다.

이득 곡선 g와 잡음 모델 λ_v인 $p_\lambda(z \mid g, \lambda_v)$의 근엽 최대화 (local maximization)를 위한 알고리즘은 다음과 같이 요약할 수 있다:

Step-0: 파라미터 초기화
$\lambda = \{a_{ij}, B^j, \sigma_j^2, i, j = 1, ..., L\}$, $g = 1$, ε가 주어지면 $p_\lambda(z \mid g, \lambda_v, g_v)$과 $l = 0$을 구한다.

Step-1: $s_n = 1, ..., L$과 $n = 1, ..., T$일때, 사후확률 $p_\lambda(s_n \mid z, g_l)$을 계산한다.

Step-2: 음질 개선: 식 (23)-(26)에 의해 $\hat{X}_j(t)\ \hat{X}_j^T(t)$을 계산한다.

Step-3: 이득인자와 잡음 파라미터의 추정:
식 (18)-(20)을 사용해서 g_{l+1}과 λ_v를 구한다.

Step-4: 만약 $p_\lambda(z \mid g_{l+1}, \lambda_{v,l+1}) - p_\lambda(z \mid g_l, \lambda_{v,l}) \le \varepsilon$이면, $\max_{g,\lambda_v} p_\lambda(z \mid g, \lambda_v) = p_\lambda(z \mid g_{l-1}, \lambda_{v,l+1})$를 할당하고 끝낸다. 그렇지 않으면 $l \to l+1$로 만들고 Step-1으로 간다.

마지막으로, 잡음이 부가된 발성 w에 대한 결정 규칙 (decision rule)은 다음과 같이 계산할 수 있다.

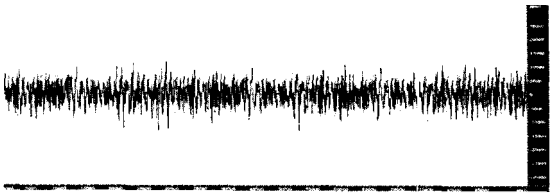
$$w = \arg \max_{1 \le i \le W} p_\lambda(z \mid W_i, g, \lambda_v) \tag{27}$$

여기서 w는 음성인식을 위한 전체 단어 수이다.

V. 실험 결과

본 논문에서는 비정상상태 자동차 잡음에 의한 왜곡된 음성 신호를 위해서 새로운 방법을 제안하였다. 음성신호의 입력 신호대 잡음비 (SNR)는 0, 5, 10, 15, 그리고 20 dB로 하였으며, SNR은 신호의 평균 전력과 잡음의 평균 전력의 비로 정의한다. 실험은 50명의 남자가 각각 고립 단어 한국어 숫자 10개 (영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)

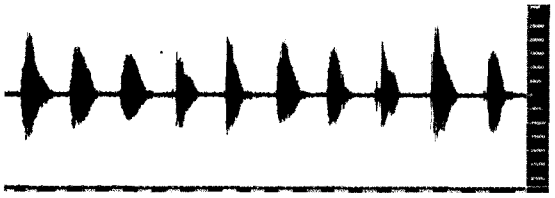
(a) 음성 신호
(a) Clean speech signal



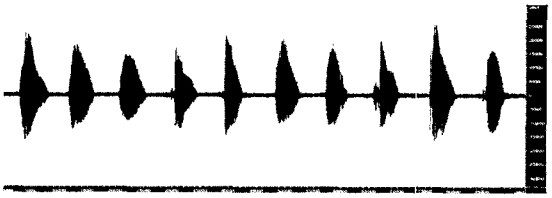
(b) 잡음 신호
(b) Noise signal

(c) 제안한 방법을 사용한 음성 파형 (SNR=0dB)
(c) speech enhanced with the proposed method (SNR=0dB)

(d) 제안한 방법을 사용한 음성 파형 (SNR=5dB)
(d) speech enhanced with the proposed method (SNR=5dB)



(e) 제안한 방법을 사용한 음성 파형 (SNR=10dB)
(e) speech enhanced with the proposed method (SNR=10dB)



(f) 제안한 방법을 사용한 음성 파형 (SNR=15dB)
(f) speech enhanced with the proposed method (SNR=15dB)

(g) 제안한 방법을 사용한 음성 파형 (SNR=20dB)
(g) speech enhanced with the proposed method (SNR=20dB)

그림 1. 제안한 방법을 사용했을 때의 향상된 신호 결과
Fig. 1. Results of speech enhanced with the proposed method.

를 3번씩 발성을 하였다. 이들 음성 데이터 중에서 30명의 화자로부터 900개의 음성데이터는 학습 과정에 사용하였고, 나머지 600개의 음성데이터는 테스트에 사용하였다. 각각의 발성은 11 kHz에서 샘플링하였으며, 상태 L=5와 15차인 AR로 모델링하였다. 학습과 인식에서 프레임 길이 N=256 샘플이고 중첩없이 사용하였다.

실험에서는 부가된 가우시안 유색 잡음의 영향에 대하여 연구하였으며, 잡음은 1500 cc 자동차에서 시속 50 Km에서 얻었다. 잡음 처리에서 잡음은 8차 AR 모델이고, 일반적으로 비정상상태 잡음에서 다중-상태 ARHMM 모델[17,18]로 확장하였다. 그림 1은 숫자음 (영~구)에 대해서 제안한 방법을 이용했을 때, 각 SNR (0, 5, 10, 15, 20 dB)에서의 음질향상된 음성파형을 나타낸 것이다.

표 1과 같이 잡음 추정없이 접근하는 방법과 제안한 잡음 추정을 하는 이득 적응된 인식 방법을 비교하였다. 잡음추정을 하는 NAR-HMM에서 제안한 Kalman 필터와 일반적인 Wiener 필터를 사용하여 결과를 비교하였다. 잡음 추정 알고리즘을 이용한 두 필터는 약 4회 반복만으로 수렴을 하였다. 표 1의 결과로부터, Wiener 필터와 Kalman 필터의 결과에서 성능의 큰 차이는 없었다. 그러나 잡음 추정을 하는 NAR-HMM은 잡음 추정이 없는 NAR-HMM보다 더 좋은 성능을 보였다. 성능향상은 신호대 잡음비가 낮은 경우에 더 두드러졌다. 제안한 방법의 잡음 추정은 오염된 음성에서 깨끗한 음성신호를 얻을 수 있도록 효과적으로 잡음을 제거하였다. 표 2는 잡음 추정을 하는 일반적인 ARHMM과 제안한 방법을 비교한

표 1. 잡음 추정을 경우와 잡음 추정을 하지 않는 NAR-HMM에 잡음이 부가된 음성인식 결과 비교

Table 1. Compared Recognition results for Noisy Speech between NAR-HMM with Noise Estimation and that without Noise Estimation.

SNR (dB)	잡음 추정을 안함 (%)	잡음 추정을 함 (%)	
		Wiener 필터	*Kalman* 필터
0	15.4	86.4	87.7
5	45.3	88.8	91.3
10	70.4	92.5	93.5
15	75.1	94.3	95.4
20	85.3	96.8	97.3

표 2. 일반적인 AR-HMM(M=0)과 제안한 NAR-HMM (M=1)의 오염된 음성의 인식 결과 비교

Table 2. Compared Recognition results for Noisy Speech between the Conventional AR-HMM (M=0) and the Proposed NAR-HMM (M=1).

SNR (dB)	*M*=0 (%)		*M*=1 (%)
	Wiener 필터	*Kalman* 필터	
0	83.6	84.5	87.7
5	86.5	87.1	91.3
10	89.4	90.0	93.5
15	93.0	93.2	95.4
20	94.7	95.1	97.3

결과이다. 잡음 추정을 하는 일반적인 ARHMM보다 제안한 NAR-HMM이 약 2-3% 향상시켰다. *M*=0일때, 제안한 NAR-HMM 방법이 일반적인 ARHMM 방법의 특별한 경우로 고려된다.

제안한 방법은 일반적인 방법보다 파라미터 수가 2배이므로 많은 계산량을 필요로 한다. 이것은 학습 데이터 크기가 클때, 요구되는 계산 시간이 길어짐을 의미한다. 그러나 증가된 계산량은 학습시간이 오프라인상에서 학습에 필요한 중요한 요소가 아니기 때문에 큰 결점으로 나타나지는 않는다. 심지어 온라인 적응 (adaptation)을 위한 증가된 계산량은 최근의 고속 시스템 속도 때문에 무시할 수 있다.

VI. 결 론

잡음이 섞인 신호만이 주어졌을 때, 이득 적응된 음성인식 (gain-adapted speech recognition)을 시간 영역에서 제안하였다. 잡음은 유색잡음으로 가정하였으며, 깨끗한 음성을 모델링하기 위해서 NAR-HMM을 사용하였다. 비정상상태 AR 처리는 알고 있는 *M*개의 기저 함수의 선형 조합으로 다항식 함수로 모델링하였다. 잡음 추정을 하는 이득-적응된 인식 알고리즘은 EM 알고리즘을 사용하여 NAR-HMM을 유도하였으며, 잡음이 부가된 음성신호의 인식에 적용하였다. 또한 다중 Kalman 필터는 음질향상 (Speech Enhancement)을 위해서 적용하였다. Kalman 필터는 각 프레임 단위로 음성과 잡음을 추정하였으며, 부가된 정상상태 유색 잡음에 대한 실험 결과는 제안한 방법이 효과적인 인식 알고리즘임을 확인할 수 있었다.

감사의 글

본 연구는 2000년 숭실대학교 교내연구비에 의하여 지원된 과제입니다

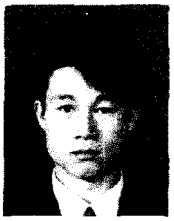
참 고 문 헌

1. J. C. Junqua and J. P. Haton, *Robustness in automatic speech recognition, Fundamentals and applications*, Kluwer Academic Pub., 1996.
2. B. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 1404-1413, Dec. 1986.
3. Y. Ephraim, "Gain adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1303-1316, June 1992.
4. L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for speech signal," *Signal Processing*, 27, pp. 65-72, 1992.
5. L. Deng, M. Aksmanovic, X. Sun, and C. F. JeffWu, "Speech recognition using HMM with polynomial regression functions as nonstationary states," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 507-520, Oct. 1994.
6. K. Y. Lee and J. Lee, "A nonstationary autoregressive HMM with gain adaptation for speech recognition," *Proc. ICSLP '98*, vol. 2, pp. 353-356, Dec. 1998.
7. L. Deng, "A stochastic model of speech incorporating hierarchical nonstationarity," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 471-475, Oct. 1993.
8. L. Deng and C. Rathinavalu, "A Markov model containing state-conditioned second-order nonstationary: Application to speech recognition," *Computer Speech and Language*, vol. 9, no. 1, pp. 63-86, Jan. 1995.
9. H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter model: Parameter selection and sensitivity to power normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 80-91, Jan. 1994.
10. M. J. F. Gales and S. J. Young, "PMC for speech recognition in additive and convolutional noise," *Technical Report CUED /F-INFENG/TR135*, 1993.

11. Y. Grenier, "Time-Dependent ARMA Modeling of Nonstationary Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. **31**, no. 4, pp. 899-911, Aug. 1983.
12. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. B*, vol. **39**, no.1, pp. 1-38, 1977.
13. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. **41**, pp. 164-171, 1970.
14. K. Y. Lee and J. Rheem, "A nonstationary autoregressive HMM and its application to speech enhancement," *Proc. Eurospeech'97*, vol. **4**, pp. 1407-1411, Sep. 1997.
15. J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. on Signal Processing*, vol. **39**, no. 8, pp. 1732-1742, Aug. 1991.
16. W. Wu and P. Chen, "Subband Kalman filtering for speech enhancement," *IEEE Trans. on Circuits and Systems*, vol. **45**, no. 8, pp. 1072-1083, Aug. 1998.
17. J. B. Kim, K. Y. Lee, and C. W. Lee, "On the application of the interacting multiple model algorithm for enhancing noisy speech," *IEEE Trans. on Speech and Audio Processing*, accepted for publication.
18. Y. Cohen, A. Erell, and Y. Bistritz, "Enhancement of connected words in an extremely noisy environment," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 2, pp. 141-148, Mar. 1997.

저자 약력

● 서 창 우 (Changwoo Seo)



1996년 2월: 창원대학교 전자공학과 (공학사)
1998년 2월: 창원대학교 전기전자제어공학부 (석사)
1999년 3월~현재: 숭실대학교 정보통신전자공학부 박사과정
※ 주관심분야: 화자인식, 음성인식, 신경망

● 이 주 헌 (Joohun Lee)

1988년 2월: 서울대학교 전자공학과 (공학사)
1990년 2월: 서울대학교 전자공학과 (석사)
1995년 2월: 서울대학교 전자공학과 (박사)
1997년~현재: 동아방송대 인터넷 방송과 조교수
2000년~현재: JSPS Postdoctoral
※ 주관심분야: 생체인식, 영상처리, 웹기반 신호처리

● 이 기 용 (Ki Yong Lee)

1991년 2월: 서울대학교 전자공학과 (박사)
1991년 9월~1997년 8월: 국립창원대학교 조교수
1994년 8월~1995년 6월: 일본 와세다대학 초빙연구원
1996년 1월~3월: 영국 에딘버러대학 박사후과정
1997년 6월~ 8월: 독일뮌헨공대 초빙연구원
1997년 9월~현재: 숭실대학교 정보통신전자공학부 부교수
※ 주관심분야: 음성신호 향상, 화자인식, 음성인식